基于深度学习的面部表情识别研究 *

陆嘉慧，张树美，赵俊莉

(青岛大学 数据科学与软件工程学院，山东 青岛 266071)

摘 要：近几年来，深度学习在语音识别、图像理解等许多应用领域取得了突破性成果。针对基于深度学习的静态人脸图像表情识别方法进行研究，首先介绍了深度学习的原理，并归纳了目前公开且常用的面部表情数据集；然后介绍了基于深度学习的表情识别的三个步骤，归纳了图像预处理和表情分类的主要方法，重点总结了目前性能较好用来提取特征的深度学习框架以及这些方法的基本原理和优劣势比较；最后指出了目前面部表情识别存在的问题和未来可能的发展趋势。

关键词：深度学习；表情识别；神经网络

中图分类号：TP391.4 doi: 10.19734/j.issn.1001-3695.2018.10.0723

Static face image expression recognition method based on deep learning

Lu Jiahui, Zhang Shumei, Zhao Junli

(*School of Data Science & Software Engineering*, *Qingdao University*, *Qingdao Shandong* 266071, *China*)

Abstract: In recent years, deep learning has achieved breakthrough results in many application fields such as speech recognition and image understanding. This paper reviews the static face image expression recognition method based on deep learning. Firstly, it introduces the principle of deep learning, and summarizes the current public and commonly used facial expression data sets. Then introduces the three steps of expression recognition based on deep learning. The main methods of image preprocessing and expression classification are summarized. This paper mainly summarizes the deep learning frameworks that are used to extract features, the basic principles, advantages and disadvantages of these methods. Finally, the problems of facial expression recognition and possible future development trends are pointed out.

Key words: deep learning; expression recognition; neutral network

0 引言

面部表情在人们日常交往的情感表达中扮演着重要角色，是识别人类的情感和行为最重要的线索之一，它被定义为对应人的内心情绪状态、意图或社交信息的脸部变化。早在 20 世纪，Ekman[1]根据跨文化研究定义了六种基本表情，这些典型的面部表情是愤怒、厌恶、恐惧、快乐、悲伤和惊喜，蔑视后来也被添加为表情之一。人脸表情识别（facial expression recognition，FER）具有广泛的应用，如人机界面、互动游戏、在线/远程教育、刑事调查和商业分析等。人脸表情识别问题是计算机视觉领域的一个传统问题，它作为智能化人机交互（human-computer interaction，HCI）技术中的一个重要组成部分，近年来得到了广泛的关注，涌现出许多新的方法。

根据不同的输入资源，面部表情识别系统可以分为两个主要的类型，即输入静态图像和动态图像序列。静态图像的方法仅从当前输入中提取特征图像，而图像序列的方法可以提取图像序列的时间信息以及每个静态图像的特征。据研究表明，传统手工提取的特征无法解决与面部表情无关的各种因素，为了解决这一问题，伴随着芯片处理能力的显著提高和精心设计的网络架构各个领域的研究已经转向深度学习方法，且这些方法已经达到了不错的识别精度。本文重点阐述基于深度学习的静态人脸图像表情识别方法。

1 深度学习

深度学习是指多层神经网络上运用各种机器学习算法解决图像、文本等各种问题的算法集合。深度学习允许由多个处理层组成的计算模型来学习具有多个抽象级别的数据表示，它通过使用反向传播算法来发现大型数据集中的复杂结构，以指示机器应如何更改其内部参数，根据前一图层的表示计算每个图层中的表示。其核心是特征学习，旨在通过分层网络获取分层次的特征信息，从而解决以往需要人工提取特征的重要难题。深度神经网络（deep neural network，DNN）训练也被称为深度学习，由于新兴强大的并行处理硬件和图形处理单元（GPU），DNN 成为模式识别和机器学习科学领域的热门话题。

深度学习的概念由 Hinton 等人[2]于 2006 年提出，表明“深度信念网络”可使用一种称为非监督“贪心逐层预训练”算法来解决深层结构相关的优化难题，之后提出多层自动编码器深层结构。此外 Lecun 等人提出的卷积神经网络是第一个真正多层结构学习算法，它利用空间相对关系减少参数数目以提高训练性能。

深度学习也为科学作出了贡献。除了在图像识别、语音识别等领域打破了纪录，它还在另外的领域击败了其他机器

收稿日期：2018-10-31；**修回日期：**2018-12-15 **基金项目：**中国博士后科学基金资助项目（2017M622137）；国家自然科学基金资助项目（61702293）；教育部虚拟现实应用工程研究中心基金资助项目（MEOBNUEVRA201601）

作者简介：陆嘉慧（1995-），女，山东青岛人，硕士研究生，主要研究方向为图像识别与处理、深度学习；张树美（1964-），女，山东莱西人，教授，硕导，博士，主要研究方向为时滞非线性系统的分析与控制、图像识别与处理（shumeiz@163.com）；赵俊莉（1977-），女，山西新绛人，副教授，硕导，博士，主要研究方向为计算机视觉、计算机图形学、虚拟现实．

学习技术，包括预测潜在的药物分子的活性、分析粒子加速器数据、重建大脑回路等。除此之外，深度学习在自然语言理解的各项任务中也取得了不错的成果，特别是主题分类、情感分析、自动问答和语言翻译。

深度学习是一个框架，其中包含多种算法。在静态图像的面部表情分析任务中，通常用到的算法有深度信念网络（deep belief networks，DBN）、自动编码器（auto-encoders，AE）、卷积神经网络（convolutional neural network，CNN）等。

2 面部表情数据集

在表情识别训练实验过程中，使用足够有效的标记数据进行训练是十分重要的，数据集应包括尽可能多的种群和环境变化。下面介绍几个比较常见且已公开的用于基本表情识别的数据集[3]。

CK+[4]：是用于评估 FER 系统使用最广泛的实验室控制条件下的数据集。包含 123 个受试者的 593 个视频序列。序列的持续时间从 10 到 60 帧不等，并显示从中性表情到峰值表情的转变。在这些视频中，来自 118 名受试者的 327 个序列基于面部动作编码系统被标记六个基本表情标签（愤怒、厌恶、恐惧、快乐、悲伤和惊讶）加上蔑视。

MMI[5]：是实验室控制条件下的来自 32 个受试者的 326 的序列，包含 740 张图片和 2 900 个视频，共 213 个序列用六个基本表情标记。此数据集存在主体差异，受试者非均匀执行相同表情，且许多受试者佩戴眼镜或留胡子等。

JAFFE [6]：日本女性面部表情数据集是实验室控制的来自 10 名日本女性的 213 个表情样本。每个主体/表情包含的样本很少。图像用六个基本表情和中性表情标记。

TFD[7]：多伦多人脸数据集是几个面部表情数据集的合并，包含实验室控制下的 112 234 张图像，其中 4 178 张用六个基本表情标签和中性表情标签注释。

FER2013[8]：是在 ICML 2013 挑战赛中引入，由 Google 图像搜索 API 自动收集的大规模且无约束的数据集，包含 28 709 个训练图像、3 589 个验证图像和 3 589 个具有六个基本表情加中性表情标签的测试图像。

AFEW[9]：自从 2013 年 EmotiW 系列情感识别挑战赛以来使用，包含从不同电影收集的视频剪辑，具有自发的表情、不同头部姿势、遮挡和照明。样本标有六种基本表情标签加中性表情。此数据集在不断更新中，2017 年 EmotiW 最新的 AFEW 7.0 包含 1 809 个视频。

SFEW[10]：是基于面部点聚类计算关键帧从 AFEW 数据集中选择静态帧创建。最常用的版本 SFEW 2.0 是 2015 年 EmotiW 的基准数据集，包含 1 766 张图片，标有六种基本表情标签加中性表情。

Multi-PIE[11]：包含实验室控制下来自 33 个视点下的 337 个受试者的 755 370 个图像。每个面部图像都标有六种表情之一：厌恶、中性、尖叫、微笑、眯眼和惊讶。此数据集通常用于多视图面部表情分析。

BU-3DFE[12]：包含实验室控制下从 100 人中捕获的 606 个面部表情序列，共 2 500 张图片，标有六种基本表情标签加中性表情。此数据集通常用于多视图 3D 面部表情分析。

Oulu-CASIA[13]：包括实验室控制下从 80 个受试者中收集的 2 880 个图像序列，标有六个基本表情标签。

RaFD[14]：包含实验室控制下 67 个受试者的 1 608 个图像，具有三个不同的注视方向，即前、左和右，标有六种基本表情标签加中性表情和蔑视表情。

KDEF[15]：是一个实验室控制下最初开发用于心理学和医学的数据集，由 70 个演员的图像组成，有五个不同的角度，包含 4 900 张图片，样本标有六个基本表情加中性表情。

3 基于深度学习的面部表情识别

通常面部表情识别可以被建模为图像分类问题，它由图像预处理、特征提取和分类三个主要步骤组成。下面简单总结了每个步骤中广泛使用的算法。

3.1 面部表情图像预处理

预处理是在提取特征之前排除与面部表情无关的一切干扰，如光照、头部姿势以及不同的背景等，目的是将面部对准到公共参考系，使得从每个面提取的特征对应于相同的语义位置。其主要方法有人脸检测、人脸对齐、数据增强、人脸归一化。

3.1.1 人脸检测

第一步是检测面部，去除背景和非面部区域。传统的人脸检测方法是利用人工提取特征来训练分类器进行人脸检测，例如 opencv 源码中自带的人脸检测器就是利用 Haar 特征进行的，但在环境变化强烈的时候检测效果不理想。Viola-Jones (V&J)人脸检测器[16]是一种经典且广泛采用的方法，已公开使用且计算简单。后来在深度学习阶段，提出了性能更好的 Faster-RCNN、R-FCN 系列以及速度更快的 YOLO、SSD 系列来检测人脸，可以适应环境变化和人脸不全等问题，但是时间久，于是又有了级联结构的卷积神经网络，进一步提高了人脸检测性能。

3.1.2 人脸对齐

虽然面部检测是实现特征学习的必要过程，但进一步的人脸对齐可以大大提高面部表情识别性能。人脸对齐可以看做在一张人脸图像搜索人脸预先定义的点，也称为人脸形状，通常从一个粗估计的形状开始，然后通过迭代来细化形状的估计。在搜索的过程中使用了两种不同的信息，即人脸的外观和形状。 形状提供一个搜索空间上的约束。广泛使用的方法是通过 IntraFace 软件，应用基于回归的面部标志定位方法，即监督下降法（SDM[18]），检测出 49 个准确的面部标志点。其他的方法有混合树结构模型（mixtures of trees，Mot[19]）、判别响应图拟合（discriminative response map Fitting，DRMF[20]）、Dlib C++库[21]、多任务级联卷积神经网络（MTCNN[22]）、DenseReg[23]和小人脸检测[24]。

3.1.3 数据增强

深度神经网络需要足够有效的训练数据以确保识别任务的普遍性，但是公开提供的 FER 数据集没有足够数量的图像用于训练，数据量少往往会导致过拟合现象。因此，数据增强是面部表情识别的关键步骤。常用的数据增强方法有旋转/反射变换、翻转变换、缩放变换、平移变换、尺度变换、对比度变换、噪声扰动、颜色变化等。同时，还有其他的如生成对抗网络生成脸，3D 卷积神经网络辅助动作单元（AUs）生成表情[25]等。

3.1.4 人脸归一化

照明和头部姿势的变化会很大程度影响面部表情识别性能，因此引入两种典型的人脸归一化方法来改善这些变化：灰度归一化和几何归一化。灰度归一化，是增加图像的亮度，使图像的细节更加清楚，以减弱光线和光照强度的影响；除了亮度调整外，还包含了对比度调整。常见的方法有直方图均衡化、基于各向同性扩散（IS）归一化、基于离散余弦变换（DCT）归一化、高斯（DoG）归一化。其中直方图均衡化效果相对最稳定，适应各种网络模型。几何归一化用来产

生正面面部视图，目前大多数还是在小角度内利用标志点对齐，最近提出了一系列基于生成式对抗网络的深度模型用于正面视图合成，如 FF-GAN[26]、TP-GAN[27]、DR-GAN[28]等。

3.2 特征提取的深度框架

基于手工提取特征的 FER 方法要手动提取与表情变化有关的有效特征，表现出有限的识别性能。随着情感识别在野外挑战（EmotiW）中的发展[30]，基于深度学习的面部表情识别问题成为了一个热门的研究课题，其关键是准确地提取每个组件。由于标注的数据实际上是有限的，所以学习模型至关重要。以下是几种近几年性能较好的基于深度学习的 FER 框架。表 1 分别从学习方式、优缺点、网络组成和典型改进四个主要方面进行了总结。

表 1 用于静态图像的表情识别深度框架比较

Table 1 Comparison of expression recognition depth frames for static images

深度学习方法	深度信念网络	自动编码器方法	深度卷积神经网络
学习方式	无监督学习	无监督学习	监督学习
优点	识别特征，分类数据，生成数据。	数据去噪；进行可视化降维；生成数据。	通过局部感知和权值共享来减少参数，可直接输入原始图像。
缺点	学习过程慢；容易导致学习收敛于局部最优解。	信息受损，数据丢失；没有全局优化。	需要调参，需要大样本量，训练时间久。
网络组成	受限玻尔兹曼机层堆叠。	神经网络编码器和解码器。	卷积层、池化层、全连接层。
典型改进	卷积 DBN、条件 RBM 等。	稀疏自动编码器、降噪自动编码器、收缩自动编码器等。	AlexNet，GoogleNet，VGGNet，ResNet，GoogleNet-Inception-Like 网络改进系列等。

3.2.1 深度信念网络

2006 年， Hinton 等人[2]提出深度信念网络及其高效的学习算法，并发表于《Science》上，成为其后深度学习算法的主要框架。训练过程包括预训练和调优过程，其中预训练过程相当于逐层训练每一个 RBM，经过预训练的 DBN 已经可用于模拟训练数据。而为了进一步提高网络的判别性能，微调过程利用标签数据通过后向传播（back propagation，BP）算法对网络参数进行微调。DBN 根据人脑分级信息处理抽象出多种表征并学习对象的特征（知识）层次，实现对知识的分层次理解。它是一种生成模型，通过训练其神经元间的权重，可以让整个神经网络按照最大概率来生成训练数据。所以不仅可以使用 DBN 识别特征和分类数据，还可以用它来生成数据。

传统的 DBN 由一堆受限玻尔兹曼机（restricted Boltzmann machine，RBM[31]）构成。RBM 由 Hinton 和 Sejnowski 于 1986 年提出，具有两层结构，即可见层和隐层，它是一个随机生成的神经网络，用于学习输入数据的概率分布，上一层 RBM 的隐层作为下一层 RBM 的可见层。随着对学习方法的进一步研究，可用于降维的 RBM 还具有很强的表示数据特征的能力，并且通常用于构建深度置信网络。Liu 等人[32]提出了一种称为增强深度置信网络（boosted deep belief network，BDBN）的方法，这是第一次系统地将特征学习、特征选择和分类器构建统一在一个框架中，BDBN 框架由一组 DBN 结构组成，每个 DBN 结构都是一个多层图形模型。BDBN 学习包括两个相互关联的学习过程，通过交替迭代直至收敛来增强分类能力。在文献[33]中提出了一个由三个连续的模块组成的行动单元（AU）启发深度网络，其中在最后一个模块中利用多层 RBM 学习分层特征。

3.2.2 自动编码器方法

自动编码器最开始作为一种数据的压缩方法，只能压缩与训练数据相似的数据，跟数据相关程度很高，且在降维的过程中会不可避免地丢失信息导致压缩后数据受损。使用自动编码器的通用神经网络通过将输出值限制为等于输入值来训练网络，使用重建输入时产生的误差来调整神经网络的每层权重。在文献[34]中介绍了深度自动编码器（DAE）来学习有效编码，与之前提到的训练用于预测目标值的网络相比，DAE 被优化以通过最小化重建误差来重建其输入。

自动编码器有多种扩展，例如去噪自动编码器[35,36]尝试编码输入且尝试重做对输入的随机损坏处理操作，是一个带有噪声破坏的自动编码器，它会产生原始输入的损坏版本。此外，还可以堆叠以获得高级功能，从而实现 SDAE（stacked denoising auto-encoder）方法，每个具有一个隐藏层的去噪自动编码器都是独立训练的，因此 SDAE 的训练是分层次的；稀疏自动编码器网络[37]加入了稀疏的限制性条件，它强制学习特征表示的稀疏性，不仅提高了算法的灵活性，而且在一定程度上使得存储变得更容易；收缩自动编码器[38]增加了一个规则项，以诱导局部不变特征，抑制训练样本在所有方向上的扰动。

3.2.3 深度卷积神经网络

最近视觉对象识别任务已经越来越多地使用“深度神经网络”，它是为了提高神经网络解决大数据问题的能力而开发的技术，提供了一种基于“类大脑”结构开发的学习体系结构。该结构可以学习多级表示和抽象，从而允许算法在图像、声音和文字中找到复杂的模式。2012 年人们发现在 CNN 中使用自动编码器做逐层预训练可以训练更深层的网络，但是后来发现良好的初始化策略要比费劲的逐层预训练更有效，随后在 2014 年提出的批量归一化（batch normalization，BN）方法对深层网络的训练起到了促进作用；到 2015 年底，通过残差网络基本可以训练任意深度的神经网络。在深度神经网络中，其中被称为“卷积神经网络”的深度模型已成为研究人员研究视觉的经典方法，广泛应用于各种计算机视觉应用。

CNN 结构由卷积层、池化层和全连接层三个主要处理层构成。卷积层对输入执行卷积，在训练过程中，选择内核和偏置参数以优化网络输出的误差函数。池化层对输入图像应用非线性变换，以减少操作后的神经元数量。在两个连续的卷积层之间放置一个池化层是很常见的，该操作还可以减小单元尺寸、减少计算负荷以及防止过度拟合问题。全连接层与经典的神经网络层完全相同，其中层中的所有神经元都连接到后续层中的所有神经元，神经元由它们的输入总和乘以激活函数传递的权重而触发。

以下介绍几个经典的卷积神经网络。1) AlexNet

深度卷积神经网络起源于 2012 年的 AlexNet[39]，如图 1 所示，这个网络应用了新的激活函数整型线性单元（ReLU）和 dropout[40]机制。AlexNet 是一种基于传统卷积神经网络分层体系结构的网络，卷积层之后是最大池层和 ReLU，在层堆栈顶部有许多完全连接的层。部分卷积层分成两个组进行独立计算，有利于 GPU 并行化以及降低计算量。在 ILSVRC-2012 比赛中，其最高失误率为 15.3%，该网络也是

第一批引入“dropout”解决过度拟合问题的网络之一，这被证明是开发大型神经网络的关键。

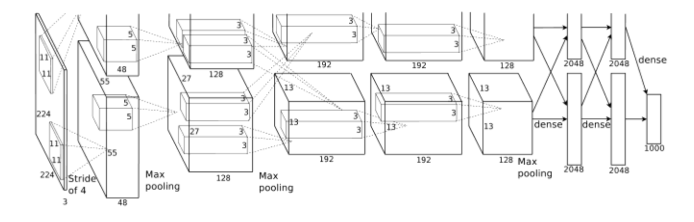


图 1　AlexNet 网络模型

Fig. 1　Alexnet network mode

CNN 架构的示意图明确显示了两个 GPU 之间的职责划分。一个 GPU 在图的顶部运行图层部分，而另一个在底部运行图层部分。GPU 只在某些层通信。网络的输入是 150 528 维的，并且网络的剩余层中的神经元数量由 253440-186624-64896-64896-43264-4096-4096-1000 给出。

2) GoogLeNet

在 2014 年 ImageNet 面向对象识别的挑战中，前三名完成者都使用了 CNN 方法，其中 GoogLeNet 架构在分类方面实现了 6.66%的显著误差率[41,42]获得第一名，它通过使用多个分类器结构，并结合多个来源进行反向传播，使用了一种新颖的多尺度方法。这种架构可以消除在到达开始层之前后向传播衰退时出现的一些问题，减少维度的附加层允许 GoogLeNet 在宽度和深度两方面面均不会有明显的损失，并且朝着 Lin 等人[43]原先描述的复杂网络网络体系结构迈出了一步。换句话说，该体系结构由多个“初始”层组成，其中每个层都像大型网络中的微型网络一样，允许架构作出更复杂的决策。该架构将多个不同尺度的卷积核和池化层进行整合，形成一个 Inception 模块，如图 2 所示，大幅度减少了模型的参数数量。

典型的 Inception 模块结构由三个三种尺寸的卷积核以及一个最大池化单元组成，它们共同接受来自前一层的输入图像，并行地对输入图像进行处理，然后将输出结果按照通道拼接起来。1*1 卷积主要用来降维。因为卷积操作接受的输入图像大小相等，而且卷积进行了 padding 操作，所以输出图像的大小也相同，可以直接按照通道进行拼接。

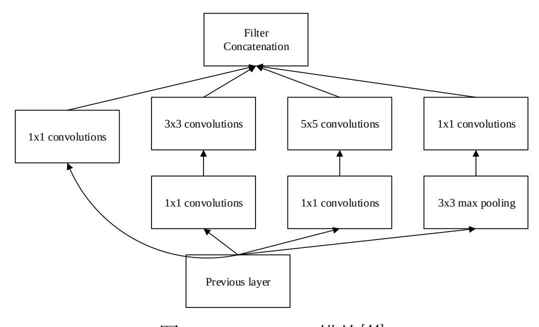


图 2　Inception 模块[44]

Fig. 2　Inception module[44]

初始深度卷积体系在文献[44]中引入，命名为 Inception-v1。后来，Inception 架构以各种方式得到了改进，首先是批量标准化的引入，对每个 mini-batch 数据的内部进行标准化处理，使每一层的输出都规范化到一个 N(0, 1)的高斯，即 Inception-v2，用两个连续的 3*3 的卷积层组成代替 Inception 模块中的 5*5 卷积层，保持感受野范围的同时又减少了参数量加速了计算，达到了 4.8%错误率。在 Inception-v3 中增加了分解思想，将一个较大的二维卷积拆成两个较小一维卷积，比如将 7*7 卷积分解成两个一维的卷积（1*7,7*1），增加了网络的非线性，达到了 3.5%错误率。Inception-v4 相较于 v3 版本增加了 Inception 模块的数量，整个网络变得更深了。

神经网络架构的改进通常依靠增加神经元数量或者增加层数，使网络学习更加复杂的功能；然而增加拓扑结构的深度和复杂性会导致一系列问题，如训练数据过度拟合以及计算需求增加。网络日益密集问题的一个自然解决方案是创建深度稀疏网络。在文献[42]中首先将 Inception 层架构应用于跨多个数据库的 FER 问题，较小的卷积应用于局部，除了从网络的稀疏性以及相对深度提供理论上的收益外，还改进了对局部特征的识别。此方法随着本地性能的提高，全局池化性能也提高了，因此不易过度拟合同时减少训练网络所需的操作次数，从而可以显著加深网络深度而不必担心使用小图像集。Burkert 等人[46]提出的网络架构受 GoogleNet 成功的启发，提出了并行特征提取模块（FeatEx），它建立了两个不同比例的并行路径，使用 1*1 大小的滤波器降低了维度，接着由 ReLU 层增强，从而创建所需的稀疏性，更好地提取了图像特征且计算量大大减少。

3)VGGNet

2014 年著名的牛津大学视觉组提出 VGG 网络[47]，并取得了 ILSVRC-2014 比赛分类任务的第二名（GoogleNet 第一名）和定位任务的第一名。这是一种只专注于构建卷积层的简单网络，结构规整，没有那么多的超参数，一个重要特性是由许多具有 3*3 小滤波器的卷积层彼此堆叠来模仿出更大的感受野的效果，而不是像先前的 CNN 模型那样使用具有更大滤波器尺寸的单个卷积层。同时 VGG 网络的拓展性很强，迁移到其他图片数据上的泛化性非常好。在 VGG 网络结构中，对图像四周各填充 1 个像素，以保证卷积后的图像大小不变。所有池化层都采用 2*2 的核，步长为 2。全连接层有 3 层，分别包括 4 096、4 096、1 000 个节点。除了最后一个全连接层之外，所有层都采用了 ReLU 激活函数。与 AlexNet 相比，VGG 去掉了 LRN 层，因为作者在实验中发现 LRN 的作用并不明显。这些思想也被用在了后续的网络架构中，如 Inception 与 ResNet。到目前为止，VGG 网络依然经常被用来提取图像特征，被广泛应用于视觉领域的各类任务。在文献[48]中对 VGG16 网络进行微调构建了一个加权混合深度神经网络（weighted mixture deep neural network，WMDNN）来自动提取特征，WMDNN 处理面部灰度图像及其相应的局部二值模式（LBP）面部图像这两个通道，输出以加权方式进行融合，LBP 和灰度人脸图像的有效结合保证了泛化能力。

4)ResNet

ResNet 网络结构的关键之处在于借鉴了“HighWay”，添加一条“捷径”连接路径。在文献[49]中提出了一个残差学习框架来减少网络的训练，它明确将层重新定义为参照层输入的学习残差函数，而不是学习未引用的函数，可以大大增加深度，提高准确度。这些残差网络的集合在 ImageNet 测试集上实现了 3.57%的 top-5 错误率，深度达 152 层，比 VGG 网络深 8 倍，但仍然具有较低的复杂度，赢得了 ILSVRC-2015 分类任务的第一名，证明了在 Inception 架构引入残差连接取得了最先进的性能，与 Inception-v3 网络相似。Szegedy 等人[41]提出具有残差连接的训练明显加速了初始网络的训练。还有一些证据表明，残差 Inception 网络的性能优于没有多余连接的 Inception 网络，提出了几个新的简化架构，即 Inception-v4、Inception-ResNet-v1 和 Inception-ResNet-v2。其中 Inception-ResNet-v1 和 Inception-ResNet-v2 是使用残差连

接的 Inception 网络，结构基本相同，只是细节不同。

另外文献[49]还分析了在 CIFAR-10 数据集上 100 层和 1 000 层的残差网络，随着网络深度的增加，神经网络的训练误差和测试误差会增大，准确度变得饱和，然后迅速退化。这种退化不是由过度拟合引起的，因为过拟合只是在测试集上的误差大。于是引入深度残差学习框架来解决退化问题，让这些图层适合残差映射，而不是每个堆叠的图层直接适合所需的底层映射，如图 3 所示，通过具有“捷径”的前馈神经网络来实现，捷径连接是那些跳过一个或多个图层的连接，只需执行标志映射，并将其输出添加到堆叠层的输出中，不会增加额外的参数和计算复杂性。整个网络仍然可以通过随机梯度下降（SGD）进行端对端的反向传播，并且可以使用通用库轻松实现，而无须修改解算器。

残差网络并不是一个单一的超深网络，而是多个网络指数级的隐式集成，在预测时，残差网络的行为类似于集成学习。对训练时的梯度流向进行分析，发现隐式集成大多由一些相对浅层的网络组成，因此，残差网络并不能解决梯度消失问题。

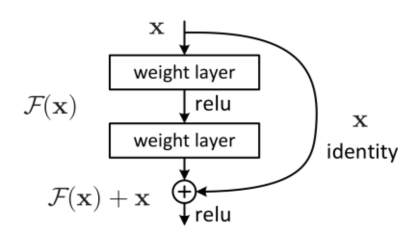


图 3　残差学习：一个构建块[49]

Fig. 3　Residual learning: building block[49]

5）其他

表 2 中列出来已应用于 FER 的一些经典 CNN 模型的网络结构和特征的比较。除了这些网络之外，还存在几个众所周知的派生框架。

在文献[50，51]中，基于区域的 CNN（region-based CNN，R-CNN[52]）用于学习 FER 的特征。在文献[53]中，更快的 R-CNN[29]通过生成高质量的区域提案来识别面部表情，避免了复杂的显式特征提取过程和低层次数据操作问题，用区域推荐网络改善了区域提案[54]质量和整体目标检测的准确性。在文献[55]中，提出了一种基于感兴趣区域（region of interest，ROI）的面部表情识别方法，将人脸图像裁剪成眼睛区域和嘴巴区域，并且针对整个脸部区域、眼睛区域和嘴部区域分别训练三个 CNN，这样不但增加了训练数据，而且不同 ROI 之间的关系有助于加强预测目标的可靠性。文献[56]中设置了七个面部 ROI 区域，使用由 Dlib 库检测的 68 个地标位置来确定面部图像的 ROI，由最小边界矩形计算，将训练数据扩展了七倍。此外在测试阶段使用了一种改进的方法，分别确定了 ROI 区域并实施决策融合策略。在文献[57]中结合感兴趣区域和 K 最近邻算法（K-nearest neighbors, KNN）提出 ROI-KNN 训练改进方法，降低了由于面部表情训练数据过少而导致深度神经网络模型泛化能力不佳的问题，从而提高了鲁棒性。文献[58]中提出了多尺度 CNN，即用两个或更多具有不同尺度输入图像的 CNN 模型，多尺度 CNN 的子 CNN 受益于各种缩放的输入图像以学习最优化参数。文献[59]中提出基于高级特征的迁移学习，即从训练的深度模型迁移高级特征来识别表情。在文献[60]中使用深度卷积神经网络特征，用于 ImageNet 对象检测的 CNN 架构被用于提取面部特征，在第五层（POOL5）上合并操作之后提取的特征被用于识别面部表情，这是特征最明显的层。Yu[61]通过使用最小化对数似然性损失和最小化铰链损失提出了两种新颖的约束优化框架来组合多个 CNN 模型，即最小化整体网络输出响应的损失来自适应的为每个网络分配不同的权重。在文献[62]中使用 Nesterov 加速梯度下降来优化的网络权重，其中参数是根据先前的动量更新，然后修正梯度以实现参数更新，这样可以防止大的振荡并提高网络的响应能力。

表 2　CNN 经典模型比较

Table 2　CNN classic model comparison

深度学习模型	时间	网络结构						识别率
		层数	卷积核大小	Dropout	数据增强	Inception	BN	
AlexNet[39]	2012	8	11,5,3	√	√	×	×	ILSVRC 2012 比赛第 1 名，最高失误率为 15.3%。
GoogLeNet[44]	2014	22	7,1,3,5	√	√	√	×	ILSVRC 2014 比赛分类任务的第 1 名，top-5 错误率 6.67%。
VGGNet[47]	2014	16/19	3	√	√	×	×	ILSVRC 2014 比赛分类任务的第 2 名和定位任务的第 1 名。
ResNet[49]	2015	152	7,1,3,5	√	√	×	√	ILSVRC 2015 比赛分类任务的第 1 名，在 ImageNet 测试集上实现了 3.57%的 top-5 错误率。

Connie 等人[63]将 SIFT（scale-invariant feature transform）特征与从原始图像学习到的 CNN 特征合并来提高识别性能，SIFT 特征用于提高小数据的性能，因为 SIFT 不需要大量的训练数据来生成有用的特征。在文献[64]中提出了一种 CNN 特征和地标特征组合的三维人脸表情识别算法，该算法由于仅使用 3D 几何脸部模型而没有任何纹理信息，所以不会出现姿态和光照变化，通过使用正交投影从 3D 人脸模型生成深度和曲率图，这些图与地标生成的组合以训练 CNN 模型。在文献[65]中主要用加权中心回归自适应特征映射（W-CR-AFM）将测试样本的特征分布转换为训练样本的特征分布，预测标签可以得到纠正，自适应特征映射可以重新构造没有标签信息的新样本的特征，从而可以纠正一些错误分类的样本。对于多视角的面部表情识别问题，Zhang 等人[66]通过在 2D SIFT 特征矩阵内对不同的面部标志点进行加权而不需要面部姿态估计，在 CNN 中引入了投影层来学习特征，以自适应地学习空间判别信息以及提取更鲁棒的高层功能，且大大减少空间复杂度，非正面面部表情识别效果较好。

3.3　面部表情分类

在学习了特征之后，面部表情识别的最后一步是将给定的图片分类输出为基本表情之一。可以将损失层添加到网络末端以调节反向传播误差，那么每个样本的预测概率可以直接由网络输出，也可以使用深度学习网络（特别是 CNN）作为特征提取工具，然后用其他分类器，如支持向量机、K 最邻近学习算法（KNN）、随机森林等进行分类。

4　结束语

面部表情识别问题一直以来是计算机视觉、模式识别领域的研究热点，尽管深度学习方法具有强大的特征学习能力，但应用于面部表情识别时仍存在问题，其鲁棒性有待进一步提高。

首先，鉴于面部表情识别是一项数据驱动的任务，训练足够深的神经网络以捕获与表情相关的细微变化需要大量的训练数据来避免过拟合，然而现有的数据集不足以训练具有深度结构的神经网络以达到最佳识别率。由于不同的年龄、种族、性别的人以不同的方式展现面部表情，所以理想的面部表情数据

集应包括具有精确面部属性标签的丰富样本图像，不仅仅只有表情标签，还应包含其他属性，如年龄、性别和种族等，这有助于使用深度学习方法对跨年龄、跨文化面部表情识别等问题进行研究。

此外，由于人脸是非刚性的形体，人脸的外观会受到成像姿势、物体遮挡、光照变化等因素的影响，这些因素与面部表情非线性耦合，所以需要深度神经网络更有效地学习特征，比如：a）可以使用网络集成方法，在特征或决策层面集成各种网络以结合它们的优势，研究表明多个网络的集合可以胜过单个网络，但在实施网络集合时应考虑两个关键因素，即网络的充分多样性以确保互补性、能够有效集合网络的适当方法；b）使用多任务网络方法，联合训练多个网络，同时考虑目标 FER 任务与其他次要任务之间的交互，许多现有的 FER 网络专注于单一任务，而不考虑其他潜在因素之间的相互作用，然而在现实世界中 FER 与各种因素交织在一起；c）使用级联网络方法，处理不同任务的各模块被顺序组合以设计更深的网络，其中前一模块的输出被后一模块利用，以分层方法顺序地训练多个网络来不断增强学习特征能力。通常这些方法可以缓解过度拟合问题，同时逐步消除与面部表情无关的因素。

与静态人脸表情图像识别相比，深度学习在视频表情分类中的应用还远未成熟。用于视频的面部表情分析可以从包含细微外观变化的动态图像序列的连续帧的时间相关性中受益，但训练的深度模型的计算量也会大大增加，与此同时训练数据的规模也在迅速增加。在未来的研究中，可以开发新的并行计算系统更加有效地利用大数据训练更大更深的深度学习模型。

除此之外，虽然微表情[17,45]不能准确识别情绪，但是可以相当准确的识别情感，也就是说通过面部肌肉的轻微活动是完全可以判断一个人是积极的状态还是消极的状态，是激动还是冷静的。如果想获得一个人的情绪应该用哪个词来形容，不仅仅需要表情和反映（以及其他唤醒水平的生理指标），还需要背景信息和他的个人经历综合起来进行理解。这种微妙而复杂的微观表情也是未来应该研究的一个方向。

参考文献：

[1] Ekman P. Facial expression and emotion [J]. Am Psychol, 1993, 48 (4): 384-392.

[2] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18 (7): 1527-1554.

[3] 徐琳琳，张树美，赵俊莉．基于图像的面部表情识别方法综述 [J]. 计算机应用, 2017 (12): 3509-3516. (Xu Linlin, Zhang Shumei, Zhao Junli. A summary of face expression recognition methods based on image, [J]. Application of Computers, 2017 (12): 3509-3516.)

[4] Lucey P, Cohn J F, Kanade T, *et al.* The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression [C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition .2010: 94-101.

[5] Pantic M, Valstar M, Rademaker R, *et al.* Web-based database for facial expression analysis [C]// Proc of IEEE International Conference on Multimedia & Expo. 2005: 317-321.

[6] Zheng Wenming, Zhou Xiaoyan, Zou Cairong, *et al.* Facial expression recognition using kernel canonical correlation analysis (KCCA) [J]. IEEE Trans on Neural Networks, 2006, 17 (1): 233.

[7] Susskind J, Anderson A, Hinton G E. The toronto face dataset, Technical Report UTML TR 2010-001, U. Toronto[R].2010.

[8] Goodfellow I J, Erhan D, Carrier P L, *et al.* Challenges in representation learning: a report on three machine learning contests [M]// Neural Information Processing. Berlin :Springer, 2013: 117-124.

[9] Dhall A, Goecke R, Lucey S, *et al.* Collecting large, richly annotated facial-expression databases from movies [J]. IEEE Multimedia, 2012, 19 (3): 34-41.

[10] Dhall A, Goecke R, Lucey S, *et al.* Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark [C]// Proc of IEEE International Conference on Computer Vision Workshops. 2011: 2106-2112.

[11] Gross R, Matthews I, Cohn J, *et al.* Multi-PIE [J]. Image & Vision Computing, 2010, 28 (5): 807-813.

[12] Yin Lijun, Wei Xiaozhou, Sun Yi, *et al.* A 3d facial expression database for facial behavior research [C]// Proc of the 7th International Conference on Automatic Face and Gesture Recognition . UK: University of Southampton, 2006: 211-216.

[13] Zhao Guoying, Huang Xiaohua, Li Stan Z, *et al.* Facial expression recognition from near-infrared videos [C]// Proc of International Conference on Pattern Recognition. 2011: 607-619.

[14] Langner O, Dotsch R, Bijlstra G, *et al.* Presentation and validation of the radboud faces database [J]. Cognition and Emotion, 2010, 24 (8): 1377-1388.

[15] Goeleven E, Raedt R D, Leyman L, *et al.* The karolinska directed emotional faces: a validation study [J]. Cognition and Emotion, 2008, 22 (6): 1094-1118.

[16] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features [C]// Proc of IEEE Computer Society Conference on Computer Vision & Pattern Recognition. USA: IEEE Computer Society, 2001: 511.

[17] 徐峰，张军平．人脸微表情识别综述 [J]. 自动化学报, 2017, 43 (3) . (Xu Feng, Zhang Junping. Facial microexpression recognition: a survey [J]. Acta Automatica Sinica, 2017, 43 (3) .)

[18] Xiong Xuehan, De La Fernando T. Supervised descent method and its applications to face alignment [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2013: 532-539.

[19] Ramanan D, Zhu Xiangxin. Face detection, pose estimation, and landmark localization in the wild [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2012: 2879-2886.

[20] Asthana A, Zafeiriou S, Cheng Shiyang, *et al.* Robust discriminative response map fitting with constrained local models [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2013: 3444-3451.

[21] King D E. Dlib-ml: a machine learning toolkit [J]. Journal of Machine Learning Research, 2009, 10 (3): 1755-1758.

[22] Zhang Kaipeng, Zhang Zhanpeng, Li Zgifeng, *et al.* Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23 (10): 1499-1503.

[23] Gyuler R A, Trigeorgis G, Antonakos E, *et al.* DenseReg: Fully convolutional dense shape regression in-the-wild [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition.2017:

2614-2623.

[24] Hu Peiyun, Ramanan D. Finding tiny faces [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition.2017: 1522-1530.

[25] Abbasnejad I, Sridharan S, Nguyen D, *et al.* Using synthetic data to improve facial expression analysis with 3d convolutional networks [C]// Proc of IEEE International Conference on Computer Vision Workshop. 2018: 1609-1618.

[26] Yin Xi, Yu Xiang, Sohn K, *et al.* Towards large-pose face frontalization in the wild [C]// Proc of International Conference on Computer Vision. Venice, Italy: IEEE Press: 2017.

[27] Huang Rui, Zhang Shu, Li Tianyu, *et al.* Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis [C]// Proc of IEEE International Conference on Computer Vision. [S.l]:IEEE Computer Society, 2017: 2458-2467.

[28] Luan Guoc Tran, Yin Xi, Liu Xiaoming. Disentangled representation learning gan for pose-invariant face recognition [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2017: 1283-1292.

[29] Ren Shaoqing, He Kaiming, Girshick R, *et al.* Faster R-CNN: towards real-time object detection with region proposal networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l]:MIT Press, 2015: 91-99.

[30] Dhall A, Goecke R, Joshi J, *et al.* Emotion recognition in the wild challenge 2014: baseline, data and protocol [C]// Proc of ACM on International Conference on Multimodal Interaction. 2014: 461-466.

[31] Hinton G, Sejnowski T. Learning and relearning in Boltzmann machines [J]. Parallel Distributed Processing, 1986, 1: 45-76.

[32] Liu Ping, Han Shizhong, Meng Zibo, *et al.* Facial expression recognition via a boosted deep belief network [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2014: 1805-1812.

[33] Liu Mengyi, Li Shaoxin, Shan Shiguang, *et al.* AU-inspired deep networks for facial expression feature learning [J]. Neurocomputing, 2015, 159 (C): 126-136.

[34] Hinton G E, Salakhutdinov R. R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[35] Vincent P, Larochelle H, Lajoie I, *et al.* Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion [J]. Journal of Machine Learning Research, 2010, 11 (12): 3371-3408.

[36] 李江，冉君军，张克非. 一种基于降噪自编码器的人脸表情识别方法 [J]. 计算机应用研究, 2016, 33 (12): 3843-3846. (Li Jiang, Ran Junjun, Zhang Kefei. Method of facial expression recognition based on denoising AutoEncoders [J]. Application Research of Computers, 2016, 33 (12): 3843-3846.)

[37] Le Quoc V, Monga R, Devin M, *et al.* Building high-level features using large scale unsupervised learning [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing.2013: 8595-8598.

[38] Rifai S, Vincent P, Muller X, *et al.* Contractive auto-encoders: explicit invariance during feature extraction [C]// Proc of the 28th International Conference on Machine Learning. 2011: 833-840.

[39] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:Curran Associates Inc, 2012: 1097-1105.

[40] Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15 (1): 1929-1958.

[41] Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-ResNet and the impact of residual connections on learning [C]// Proc of Computer Vision and Pattern Recognition. 2016.

[42] Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks [C]// Proc of IEEE Winter Conference on Applications of Computer Vision .2016:1-10.

[43] Lin Min, Chen Qiang, Yan Shuicheng. Network in network [J]. Computer Science, 2013. https://arxiv. org/abs/1409. 4842.

[44] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2015: 1-9.

[45] 唐爽. 基于深度神经网络的微表情识别 [J]. 电子技术与软件工程, 2017 (3): 93-95. (Tang Shuang. Microexpression recognition based on deep neural network [J]. Electronic Technology&Software Engineering, 2017 (3): 93-95.)

[46] Burkert P, Trier F, Afzal M Z, *et al.* DeXpression: deep convolutional neural network for expression recognition [J]. Computer Vision and Pattern Recognition, 2015, 22 (10): 217-222.

[47] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014. https://arxiv. org/abs/1409. 1556v6.

[48] Yang Biao, Cao Jinmeng, Ni Rongrong, *et al.* Facial expression recognition using weighted mixture deep neural network based on double-channel facial images [J]. IEEE Access, 2017 (99): 1-1.

[49] He Kaiming, Zhang Xiangyu, Sun Jian, *et al.* Deep residual learning for image recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV:United States, 2016:770-778.

[50] Sun Bo, Li Liandong, Zhou Guoyan. Facial expression recognition in the wild based on multimodal texture features [J]. Journal of Electronic Imaging, 2016, 25 (6): 061407.

[51] Sun Bo, Li Liandong, Zhou Guoyan, *et al.* Combining multimodal features within a fusion network for emotion recognition in the wild [C]// Proc of ACM on International Conference on Multimodal Interaction. 2015: 497-502.

[52] Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE Computer Society, 2014: 580-587.

[53] Li Jiaxing, Zhang Dexiang, Zhang Jingjing, *et al.* Facial expression recognition with faster R-CNN [J]. Procedia Computer Science, 2017, 107 (C): 135-140.

[54] Uijlings J R, Sande K E, Gevers T, *et al.* Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104 (2): 154-171.

[55] Cui Ruoxuan, Liu Minyi, Liu Manhua. Facial expression recognition based on ensemble of mulitple CNNs [J]. Biometric Recognition. 2016: 511-518

[56] Sun Xiao, Lv Man, Quan Changqin, *et al.* Improved facial expression recognition method based on ROI deep convolutional neutral network [C]// Proc of International Conference on Affective Computing & Intelligent Interaction. [S.l.]:IEEE Computer Society, 2017: 256-261.

[57] 孙晓，潘汀，任福继. 基于 ROI-KNN 卷积神经网络的面部表情识别 [J]. 自动化学报, 2016, 42 (6): 883-891. (Sun Xiao, Pan Ting, Ren Fuji.

Facial expression recognition based on ROI-KNN convolutional neural network [J]. Automation journal, 2016, 42 (6): 883-891.)

[58] Zhou Shuai, Liang Yanyan, Wan Jun, *et al*. Facial expression recognition based on multi-scale CNNs [C]// Lecture Notes in Computer Science. Cham:Springer, 2016.

[59] Xu Mao, Cheng Wei, Zhao Qian, *et al.* Facial expression recognition based on transfer learning from deep convolutional networks [C]// Proc of IEEE International Conference on Natural Computation. 2016: 702-708.

[60] Mayya V, Pai R M, Manohara P M M. Automatic facial expression recognition using DCNN [J]. Procedia Computer Science, 2016, 93: 453-461.

[61] Yu Zhiding. Image based static facial expression recognition with multiple deep network learning [C]// Proc of ACM on International Conference on Multimodal Interaction. 2015: 435-442.

[62] Su Wanjuan, Chen Luefeng, Wu Min, *et al*. Nesterov accelerated gradient descent-based convolution neural network with dropout for facial expression recognition [C]// Proc of Asian Control Conference. 2017: 1063-1068.

[63] Connie T, Al-Shabi M, Cheah W P, *et al.* Facial expression recognition using a hybrid CNN-SIFT aggregator [C]// Proc of International Workshop on Multi-disciplinary Trends in Artificial Intelligence. Cham: Springer, 2017: 139-149.

[64] Yang Huiyuan, Yin Lijun. CNN based 3D facial expression recognition using masking and landmark features [C]// Proc of International Conference on Affective Computing & Intelligent Interaction. [S.l.]:IEEE Computer Society, 2017: 556-560.

[65] Wu Bingfei, Lin C H. Adaptive feature mapping for customizing deep learning based facial expression recognition model [J]. IEEE Access, 2018, 6: 12451-12461.

[66] Zhang Tong, Zheng Wenming, Cui Zhen, *et al*. A deep neural network-driven feature learning method for multi-view facial expression recognition [J]. IEEE Trans on Multimedia, 2016, 18 (12): 2528-2536.